ON SOME APPLICATIONS OF PATTERN RECOGNITION AND IMAGE PROCESSING TO CYTOLOGY, CYTOGENETICS AND HISTOLOGY

by

JUDITH M. S. PREWITT

Värmlands Nation.

Department of Computer Science, Uppsala University

and

Department of Clinical Cytology, University Hospital, Uppsala.

UPPSALA 1978

ON SOME APPLICATIONS OF PATTERN RECOGNITION AND IMAGE PROCESSING TO CYTOLOGY, CYTOGENETICS AND HISTOLOGY

by

JUDITH M. S. PREWITT

Värmlands Nation.

Doctoral dissertation to be publicly discussed in Hörsal 1, Humanistiskt Samhällsvetenskapligt Centrum, Kyrkogårdsgatan 10, Uppsala University on May 30:th, 1978 at 10 p.m. for the degree of Doctor of Philosophy.

Department of Computer Science, Uppsala University

and

Department of Clinical Cytology, University Hospital, Uppsala.

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41548924_0015

ON SOME APPLICATIONS OF PATTERN RECOGNITION AND IMAGE PROCESSING TO CYTOLOGY, CYTOGENETICS, AND HISTOLOGY

by

Judith M. S. Prewitt

Department of Computer Science, Uppsala University
and
Department of Clinical Cytology, University Hospital,
Uppsala

CONTENTS

1. INTRODUCTION 3

2. SPECIMEN PREPARATION AND DATA ACQUISITION 8

3. APPLICATIONS TO CYTOLOGY 13

4. AN APPLICATION TO CYTOGENETICS 32

5. APPLICATIONS TO HISTOLOGY 37

6. SPECIALIZED SOFTWARE 43

7. ACKNOWLEDGEMENTS 45

8. LIST OF PAPERS 46

UNIV.
LEIDEN
BIBL.

1. INTRODUCTION

Pattern recognition is a set of applied mathematics techniques which aims at characterizing and differentiating objects of interest based on descriptive features or properties and relationships of the objects. Image processing is a set of applied mathematics techniques which aims at extracting descriptive features and relationships of pictorial objects of interest as a prelude to pattern recognition. The burgeoning volume of pictorial data which has been subject to pattern recognition comes from many fields of application: earth resources, satellite weather prediction, radiology, thermography, ultrasonics, two- and three-dimensional tomographic reconstruction, fingerprinting, handwritten signatures, voice and speech recognition, robotics, assembly line inspection, and last but hardly least, microscopy.

The subject of this thesis is applications of pattern recognition and (digital) image processing to cytology and histology - the study of cells and the study of tissues based on microscopy. Four specific applications are encompassed: (1) white blood cell (leukocyte) differentiation, (2) recognition of different cell types in human breast aspirates, (3) chromosome karyotyping, and (4) quantitative characterization of tumors of the human urinary bladder epithelium. None of these (digital) image processing applications would be possible were it not for the recent advancement in the state-of-the-art of modern, large, high speed digital computers and of reliable, sensitive, precise, stable and rapid microscope field imaging devices (digital scanning microscopes).

The benefits of computer involvement in microscopy and cytology/histology automation are fivefold:
(1) Achievement of greater resolution, reliability, reproducibility and objectivity through quantitation rather than through subjective visual evaluation.
(2) Minimum error computation in feature extraction (e.g., computation of nuclear extinction, shape descriptors, measures of nuclear eccentricity, symmetry, inhomogeneity, area, nuclear to cytoplasmic ratios, etc.).
(3) Throughput unachievable by manual means on a scale suitable for valid statistical studies of populations of cells.
(4) Recording and evaluation of subtle and possibly important characteristics which may pass unnoticed during conventional cytopathological and histopathological examinations or which are not directly obtainable (perceptible).
(5) Lower costs of health care delivery.

The interpretation of pictorial data has been dominated throughout the history of science by the role of the human eye. This

is as true for the images of biological objects in microscopy as it is for most other optical applications. The trained pathologist mentally catalogues the size, shape, color and texture of microscopic fields of view with the purpose of recognizing pathologies. He is able to quickly and readily extract diagnostic information from microscope specimens by perceptual and cognitive processes which seem to involve selective attention to visual cues and unverbalized impressions in some poorly understood way. Both spontaneous observation and learning gained through repetitive general and specialized experience play roles. As proficient as the individual pathologist may become, he is limited by the inherent low classification resolution and high subjectivity of the traditional qualitative method of morphological description. The limitations of this approach and the uncertainty in interpretation which it generates is reflected by a multitude of confusing and conflicting systems of nomenclature and classification.

However, if one may judge from recent progress in automated image analysis involving other fields, the eye's dominance in microscopy may soon be challenged. Man may continue to look down microscopes, but he is already using machines to free himself from the drudgery of routine searching, sizing and counting, and he may in the distant future have automata as colleagues in the subtle interpretation of microscope images. Such hypothetical machines could revolutionize the processing of large volumes of routine microscopic specimens currently analyzed by trained technicians in laboratories and hospitals. In addition, they offer the exciting possibility of the development of an analytic prowess considerably beyond the capability of the human eye-brain.

Automation of the acquisition and interpretation of data in microscopy has been a focus of biomedical research for two decades. In spite of many serious attempts, mechanical perception of microscopic fields with a reliability that would inspire routine application is just now becoming possible.

Many facets of the problem appear to be well within the grasp of present-day technology. Thus, available cytochemical and histochemical techniques make it possible to prepare biological materials so that morphological integrity is preserved, key constituents are stained stoichiometrically, and the specimens are favorably disposed for effective imaging. Scanning microscopes now have the requisite sensitivity, resolution, stability, precision and rapidity to sample such objects and make photometric measurements over a wide range of magnifications and wavelengths of light within the visible and near-visible spectrum. Further-

more, modern large capacity, high speed automatic digital computers at last provide the ability to manipulate and process the hitherto unmanageable quantities of optical information contained within all but the simplest images. Thus, by coupling computers with optical sensors, we can anticipate a new class of machines potentially able to analyze microscopic images.

Given such machines, the problem then is to teach them image analysis. This may be in terms of human criteria, but it is probably more appropriate that machines be given a frame of reference defined by the characteristics of their own perception. One may expect developments in microscopy to parallel those in other areas of machine perception. There are already machines that read print and travel through space to observe the planets,and efforts are being made in the computer interpretation of roentgenograms and aerial photographs. The application to microscopic images has some features in common with these other types of image analysis, but there are important differences.

With the basic materials for achieving automation via mensuration at hand, attention turns toward generating and evaluating methods for extracting meaning from quantitative optical information. Definitive concepts for image structuring and image characterization have to be realized, to be given satisfactory operational definitions, and to be assembled within a machine-oriented perceptual framework.

Criteria for effective and efficient discrimination and interpretation of images have been evolved. It would have been a serious mistake to presuppose that mechanical perception must mimic the human's perceptual apparatus in organizing images as complexes of picture elements. Likewise, it would have been serious to ignore traditional descriptive morphology and established taxonomies. In steering a middle course, the exploration of many complementary approaches and the introduction of numeric methods which fully utilize measurements of optical density or light intensity seemed to hold the most promise for explicating and augmenting the existing, largely verbal tradition of microscopic morphology.

To realize our objectives, we designed a system with three basic components: (1) a sensor which accurately and effectively scans microscope fields and converts the optical information into digital form; (2) human analysts who contribute heuristics, devise image processing methods and ensembles of discrimination criteria, program and supervise the other components, systematically set standards, and select and prepare the material to be studied; and (3) a large general purpose digital computer which assimilates the enormous volumes of quantitative optical information and ren-

ders the machine interpretation of images. Preprocessing of the optical information by the sensor is kept to a minimum, and the quality of the optical information transmitted to the computer approaches the theoretical limit imposed by the microscope itself. Thus, the heuristic and algorithmic aspects of image processing can interact without restraints forced by fixed logic hardware, and a facile and free interplay is encouraged between the analyst and the computer.

The paradigm for image analysis and pattern recognition applied to microscopy is illustrated in Figure 1. The computer analysis of an image scan may be represented by a sequence of nine stages: (1) preparation, (2) data acquisition, (3) preprocessing, (4) delineation of figure and ground (global scene analysis), (5) articulation of subordinate morphological features (local scene analysis), (6) extraction of image-characterizing parameters (features, properties, relationships), (7) selection of best discriminating sets of parameters, (8) development of taxonomic systems, (9) classification of unknown specimens. Note the recursion. Any system for analyzing complex images will likely require an iterative interaction of the various steps of processing.

Preparation refers to the selection of a sample of the objects of interest from the patient, and the staining of the specimen to render its morphologic and photometric properties visible to the computer's "eye." Preprocessing refers to the initial reading of data and the translation and organization of the data into a suitable, compact format for subsequent analysis. The scan is converted into a matrix of optical density (brightness) values.

A scan may represent a single object or a complex comprising several objects of interest. It may have an empty background, a background slightly mottled by artifacts of fixing and staining, or a background replete with optical information. Delineation of figure and ground (global scene analysis) refers to the sequence of transformations and computations that isolates or demarcates or segments the individual objects of interest. Articulation of subordinate morphological features (local scene analysis) carries the process one step further and uses similar methods to subdivide the object of interest into relevant constituent parts.

After an object or its subdivisions (semantically motivated) are delineated, the stage is set for the extraction of image-characterizing parameters (or features or descriptors). The boundaries of objects themselves provide information about the morphology and topography of the image, and optical density or brightness

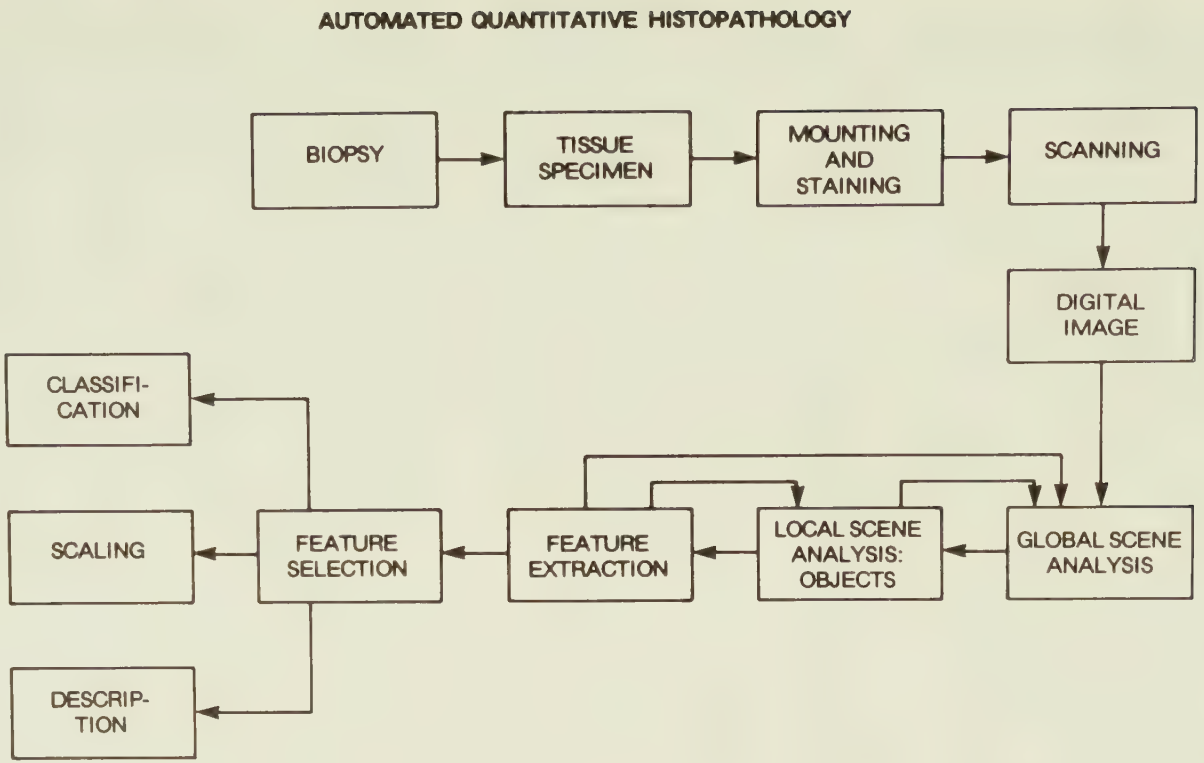

Figure 1. The image processing and pattern recognition paradigm for automated quantitative microscopy. Note the recursion.

information can be used to indicate the texture and chemical content of specific parts of the object. For the sake of parsimony, the armamentarium of numeric features that can readily be generated is honed down to a minimal descriptive set by feature selection.

The image descriptors and particularly the numeric parameters, are used for the development of taxonomic systems. The descriptors simplify the difficult task of finding similarities and differences in images by replacing it with the more tractable task of finding similarities and differences from measurements derived from these images. Images are grouped into classes on the basis of many measured attributes, using mathematical and statistical techniques such as discriminant analysis and clustering - supervised and unsupervised learning in the technical vocabulary of pattern recognition. The features themselves are grouped into classes on the basis of canonical analysis and intersubstitutability properties. Classification of unknown specimens can then be done on the basis of the similarities of extracted parameters to those of the established taxonomic classes.

2. SPECIMEN PREPARATION AND DATA ACQUISITION

The aim of specimen preparation and data acquisition is to transduce a microscopic image into a series of digits and to do this with fidelity, reasonable efficiency, and a minimum of preprocessing. This is analogous to transforming an image into silver grains in photography. To get a good photomicrograph one must begin with a good preparation; appropriately match the stains, light source and color sensitivity of the emulsion; carefully control exposure, contrast and latitude; and finally choose the magnification, resolution and other optical variables appropriate to the problem. Instead of generating a local area of blackening on the film, we want a single number representing local optical density; and instead of such areas forming a two-dimensional image, we want an array of numbers to be assembled into what is essentially a digital image in the memory banks of the computer. This is accomplished by making measurements with a scanning cytophotometer of light transmittance or absorption for a large number of points distributed generally in a Cartesian lattice over the microscopic field of view. The continuous measurements (or gray levels as they are sometimes called) are sampled and then digitized into a matrix of 4 to 8 bit bytes. From this, an array of digits or gray levels is formed and constitutes a computer-compatible image or picture, quantized both spatially and photometrically.

There are two primary types of television microscopes (digital scanning cytophotometers), image plane scanners and object plane

scanners. In the image plane scanner, the light detector occupies a plane conjugate to the eyes of the microscopist. This is easily instrumented using commercial microscopes which are designed for a trinocular tube for photomicrography. In the object plane or flying spot scanner, the illuminating source is placed in the normal image plane of the microscope so that reverse path optics image this source into the specimen. The size of the source (a beam traversing the face of a cathode ray tube in television raster pattern) minified must be small and equals that of a picture element or pixel. Hence the name, "flying spot." As the flying spot addresses the specimen, the light transmitted through the specimen is collected by the microscope condenser and impinges on a light detector.

Paper II discusses instrumentation, performance evaluation by eye and formally, spatial and photometric resolution, and some elementary data processing applications.

The CYDAC flying spot scanner, a prototype instrument, was used exclusively for the research on characterizing and differentiating white blood cells and chromosomes. CYDAC has five main subsystems: (1) a flying spot microscope for scanning the image, (2) a photometer for measuring the optical density continuously during the scan, (3) a converter for transforming the photometric measurements into digital form, (4) a magnetic tape recorder for recording measurements, and (5) operator controls and displays for monitoring the system. It has the following basic design features: spot size, 0.25 μm; magnification, up to 800x; intersample distance, 0.25 μm; gray levels, 256 linear in optical density; spectrum, wide bandpass centered at 503 nm.

The operation of the instrument can be briefly summarized as follows. Stage motion and focussing of the microscope are manually controlled, and the microscope can be used either as a conventional instrument by viewing through the eyepiece or as a television microscope by viewing a high-speed video display presented by the search monitor. The operator selects the field to be scanned, sets the photometric controls, focusses the microscope, and activates the recording process. CYDAC then automatically shifts to a slow-speed scanning mode, waits a brief interval for stabilization, records a complete scan of the chosen field, and reverts to the high-speed search mode.

Scanning is done under electronic control, using a cathode ray tube having a rectilinear TV raster scan pattern to produce a moving light source. Blue green light from the scanning spot enters the microscope through the eyepiece and is focussed by the objective onto the mounted specimen where it is reduced to 0.25 micron. The light intensity is modulated by the specimen

and is then converted into an electrical signal by a substage photomultiplier. A second photomultiplier senses light coming directly from the cathode ray tube. The former light path is called the data channel and the latter, the reference channel. The signals from the two photomultipliers undergo logarithmic transformation and their difference is then an analogue representation of optical density along the path of the scanning spot. The electrical output of the system represents the transmission or optical density. The signal is sampled periodically as the spot of light moves uniformly. It is then converted from analogue to digital (binary) form and recorded on magnetic tape as the scan proceeds. CYDAC has no facility to play back or analyze these tapes, and the subsequent processing required submitting the finished magnetic tapes to a compatible general purpose digital computer.

The optical density of the specimen is measured along lines spaced 0.25 micron apart, with the interval between successive samples on a line also fixed at 0.25 micron. Each observation contains eight bits of information, giving a gray scale quantization of 256 levels over an adjustable linear optical density scale. Spatial quantization is effected by the act of sampling the output of the moving spot. The regularly spaced sequence of samples, along with appropriately interspersed end-of-scan-line markers, suffices to coordinate each observation with its corresponding position in the object.

A 50 x 50 micron field is sensed and represented on magnetic tape in six seconds as 40,000 eight-bit optical density measurements forming a 256 x 256 pixel array. The signal-to-noise ratio is 100:1. The modulation transfer function of the system approaches the theoretical diffraction limit of optical resolution. For example, the maximum spatial frequency for 50 percent transmissivity in CYDAC is 1150 cycles per mm in the specimen plane, while the theoretical limiting frequency for the same transmissivity is 2000 cycles per mm at a wavelength of 500 nm and a numerical aparture of 1.25. The instrument has an uncertainty in incident light of 1.0 per cent CYDAC value and no measurable distortion of the scanning raster in the specimen plane.

On reading a CYDAC tape, the digital computer receives a primary representation of the scanned field as a matrix of optical density values in which spatial and gray-scale relationships are preserved. A set of graded alphanumeric symbols simulating a 32-level gray scale can be used to reconstitute CYDAC scans on the high-speed printer in the absence of other display devices. Such computer reconstitutions of the images of human blood cells, originally scanned by CYDAC at 256 levels, retain sufficient in-

formation for ready morphological discrimination by eye. As this thesis shows, the primary representations themselves contain sufficient information for morphological discrimination by machine.

Paper I is concerned with the problem of selecting and appraising a sampling interval for digital data collection systems in general, and for the CYDAC scanning cytophotometric data acquisition system in particular. A conservative point of view suggests that, for adequate reconstruction and subsequent quantitative analysis of microscope images, the scanner must provide field of view, spatial resolution and resolution of light intensity comparable to that which the human as sensor and processor can apparently utilize. Because spatial resolution and gray-scale resolution are not separable concepts, the ultimate choice of design parameters (spot size, scanning rate or spot velocity, and sampling rate) will represent a compromise.

The mathematical development is based on an ideal (noise free) exemplar signal produced by scanning a regular pattern with abrupt changes in transmission (a black and white bar pattern). An exact harmonic analysis is made of this output response to scanning a high contrast object. Finally a set of signal reconstructions is made by trigonometric synthesis. The relative residual, a figure of merit analogous to the coefficient of variation, is introduced to measure the departure of the reconstructed signal from the original signal; it is sensitive to both sampling interval and the model for signal reconstruction. With estimator $\hat{s}(t)$ of the signal $s(t)$, we associate the residual η, the mean residual σ^2, and the relative residual E, defined respectively as

$$\eta^2 = \int_{t_o}^{t_o+T} \left[s(t)-\hat{s}(t)\right]^2 dt$$

$$\sigma^2 = \eta^2/T$$

$$E = \sigma/\mu \qquad (\mu \neq 0)$$

The relative residual is an explicit function of spot diameter, spot velocity, and sampling rate. With the first two of these elements fixed largely by technological and economic considerations, a sampling rate was sought which matched the relative residual of the reconstructed ideal signal to the noise characteristics of the scanning system. The relative residual E is analogous to the coefficient of variation and its reciprocal 1/E is reminiscent of signal to noise ratio. E, η^2 and σ^2 measure the error of estimation. Any one of them can serve as

a figure of merit for assessing the joint adequacy of a method of signal reconstruction and sampling rate. Signal recovery, i.e., emulation of those aspects of the original function which are of interest, will depend on the quality and nature of the input, on the sampling interval, and on the effectiveness of the technique used for signal reconstruction. An explicit relationship between the relative residual and the sampling interval is derived for Fourier methods, since they were the most suitable of several methods of digital synthesis studied. The relative residual for best trigonometric expansion of the signal involves Bessel functions of order 1.

Cells in breast aspirates were imaged with the flying spot scanner at the Department of Clinical Cytology, University Hospital, Uppsala. Its microscope is equipped with a stage which can be moved in two orthogonal directions using two pairs of stepping motors. The first pair can move the stage in 0.5 micron steps, while the other pair can move the stage in steps of 10 microns. To enable automatic focussing, a stepping motor has been attached to one of the focussing knobs. The illumination is provided by a stabilized xenon arc lamp, either via a continuously variable grating monochromator or via narrow bandpass interference filters through a 0.7 micron square aperture.

The research on epithelial tissue sections of the human urinary bladder was carried out using the Jet Propulsion Laboratory's Image Processing Laboratory's ALMS-2 system. ALMS-2 is a high precision vibrating mirror scanner in which a mirror redirects a spot of light in sweeping fashion across the specimen. The scanned field of view is about 500μm by 500μm, with 0.5μm line spacing and spot size, and hence pictures measure as large as 1024 by 1024 pixels.

As a first study in image characterization and discrimination, we selected the normal human peripheral blood smear and concentrated our attention on a sample of 100 typical members of five classes of leukocytes (neutrophils, lymphocytes, monocytes, eosinophils, and basophils). The cells are prepared by a flotation technique which eliminated red cells and dispersed the leukocytes as isolated cells. They were then smeared on slides, air dried rapidly, and stained with gallocyanin chrome alum and naphthol yellow S, reagents that preferentially bind to nucleic acids and protein respectively.

Chromosomes were prepared in the following manner. Whole blood was cultured for three days, the last five hours of which colchicine was added. The cultured cells were placed in hypotonic solution, smeared on slides and then fixed and air dried. The preparation was digested with ribonuclease and stained with gallocyanin chrome alum for three days, and mounted in Cargille oil.

Breast cells from fine needle aspiration biopsies were prepared in the usual manner and stained with Papanicolaou stain. Sections of papillary growths of the human urinary bladder epithelium were embedded in parafin, sliced 0.4 micron thick and stained with hematoxylin to render nuclear detail.

Various computers were used in the course of these applications. Research on white blood cells and chromosomes was started on a Honeywell 800 computer, thence switched to an IBM 7040, and finally relocated to an IBM 360/65. The breast aspiration biopsy project and the epithelial tissue project were both conducted on a DEC PDP-10.

3. APPLICATIONS TO CYTOLOGY

The leukocytes of peripheral human blood are important indicators in the diagnosis and evaluation of disease, and the differential and total white cell counts have accordingly become vital routine medical laboratory procedures. As customarily performed, these tests require microscopic examination of specially stained blood smears by technicians who select an area of the microscope slide, and draw morphological and quantitative conclusions from a random sample of approximately 100 white cells. The leukocytes fall into five morphological categories- lymphocytes, monocytes, neutrophils, eosinophils, and basophils - whose distinguishing characteristics and abnormal deviations have been copiously described and illustrated in medical texts and hematological atlases. Purported visual cues include size, color, texture, granularity or local structure, relative intensity, and shape of the cells and their nuclei. Results, naturally, are limited by the peculiarities of the human visual information processing system. They are sensitive to human factors such as observer attention span, fatigue, sampling bias, self consistency and external consistency, and can therefore be both imprecise and inaccurate. 300 million blood smears are read annually in the United States at a cost of $1500 million. Supplanting these traditional visual diagnostic tests with automatic machine analysis offers the prospect of increased reliability and speed, and moreover, the possibility of recasting the qualitative and subjective basis of diagnosis into quantiative and objective terms.

The cornerstone of the automation of the white cell differential count is decision theory. Four major topics of decision theory are relevant: (1) discriminant analysis and in particular Beyesian analysis, (2) feature selection, (3) canonical analysis and (4) cluster synthesis or distribution free analysis. Papers III and IV discusses results obtained with all these methods of approach.

First to be shown was that descriptors oriented around patterns of light distribution as well as topology or shape and color can be extracted, combined into a parsimonious vectorial representation of the cells, and used in a robust linear discriminant function model for recognition. To do this, we adopted the classical approach of statistical decision theory, applied it to samples of typical leukocytes selected and identified by a cytomorphological authority and designed a reliable and efficient linear recognition logic.

Strategies of four types entered into the decision-theoretic approach: (1) strategies for classification, (2) strategies for affixing values to parameters of the decision process, (3) strategies for the selection and rejection of features or variables, and (4) strategies for the evaluation of performance. For each of these types of strategies, figures of merit and optimality criteria were needed. In order to obtain good algorithms and have confidence in predictions of their behavior, the best use must be made of available sample information. This means simultaneously optimizing the (1) number of subjects, (2) number of features or variables per subject, (3) measurement resolution of each variable, (4) apportionment of samples among the several categories in both design and test sets, and (5) complexity of the analysis - all this in the light of (i) any structural assumptions about the underlying probability distributions of subjects and variables, (ii) risks or penalties of error, (iii) the cost of data acquisition and measurement, and (iv) the cost of computation. Selection of variables for parsimonious characterization and optimal, efficient decision making, and specific discriminatory achievements are meaningful only in the context of admissible discriminant functions, the available learning samples, and the method of statistical estimation which was utilized.

Second to be shown was that the photometrically oriented descriptors permit articulation of pattern classes corresponding to the five commonly accepted leukocyte types, on the basis of clustering phenomena in the feature space alone, without invoking parametric or probabilistic assumptions about the underlying signal structure. In addressing this second problem, we devised a non-parametric cluster analysis to look for this type of consistency (Paper VI).

Third to be shown was that the photometrically oriented descriptors that make machine recognition so facile are not more obscure or abstract than the morphological and spectral characteristics on which we claim to rely. They can be assigned meaningful interpretations in natural language, and fall into descriptive families: shape, content, concentration, contrast and comparisons (ratios). Moreover computer synthesized pseudo-cells images, embodying the statistical properties of the photometrically oriented descriptors

without simultaneously incorporating shape and textural characteristics, can readily be distinguished and sorted by eye into classes corresponding to the cell types with photometric qualities which they emulate. The human observer can thus appreciate what the computer recognition accomplishes.

Leukocyte images presented for computer characterization and identification consisted of individual components, which, because of the biochemical and biomechanical preprocessing, were noncontiguous and nonoverlapping, and dispersed against a uniform and structureless background.

The initial tasks of image processing were (1) differentiation of the cell from background, and of the nucleus of the cell from the cell cytoplasm (scene segmentation and image articulation), and (2) establishment of a reference gray level on a per image basis, which would allow normalization and meaningful comparison of images and their descriptors. Both of these functions were fulfilled by a method of image articulation based on machine inspection of histograms of gray values appearing in the image and density thresholding (Papers III and IV). The novel method of implicit spatial differentiation uses data-generated relative thresholds and is called the frequency distribution or histogram method. The method is suitable for pictures in which (1) objects appear as regions of fairly constant gray value on a fairly uniform ground of different average gray value; (2) edges appear as regions of intermediate gray values and rapid density flux; (3) objects occupy a greater proportion of image area than edges; and (4) object regions are either disjoint or satisfy set-theoretic inclusion relationships. The frequency distribution method combines gradient-maximizing and thresholding by identifying those gray levels which are embedded in steep gradients. Scan elements on a steep gradient of optical density will contribute sparsely to the frequency of occurrence of optical density or gray values spanning the gradient. Accordingly, objects and background will generate histogram modes, while edges correspond to antimodes. The global threshold for delineating an object is given by the gray value at the antimode, and the background mode provides a natural baseline for anchoring the histogram and standardizing feature values.

For four of the cell types - lymphocytes, monocytes, neutrophils, and monocytes - the gray value histograms have a trimodal trend, reflecting the presence of background, cytoplasm, and nucleus. For the fifth cell type - basophils prepared by our method - the histogram is bimodal because the cytoplasm is transparent to the scanner at the wavelength of light used. In a field containing an intact leukocyte, the local minimum with the lowest optical

density corresponds to the cytoplasmic boundary and provides a threshold optical density value which is effectively an upper bound for background pixels and a lower bound for cell pixels. The local minimum with the highest optical density defines the nuclear boundary and divides the cell into cytoplasmic and nuclear elements.

Informal visual comparison of algorithmically thresholded images with both unprocessed images and spatially differentiated images indicated concurrence of visually placed boundaries, computer-drawn boundaries, and the locus of peak gradient magnitudes. For this purpose, high speed printer reconstitutions of gray scale images using a 32 character print code to simulate gray values was employed for showing both the original scan of the microscope field and for gradient magnitudes.

More formal tests which stressed analogs of our own perception of edges confirmed these informal judgments. First, examination of individual density and gradient profiles taken orthogonal to histogram-generated cell and nuclear boundaries, and composites of such profiles, showed that these boundaries coincide with the inflection point or position of maximum gradient magnitude. Second, statistical testing of serial values in density and gradient profiles indicated more significant heterogeneity near a boundary than within background or within a major cell constituent. Finally, assessment of the homogeneity of several cell types, using boundary-dependent properties which are known independently to express cell invariants preserved under imaging, provided an external test of consistency for the boundary positions. The closer the estimate of the invariant is to the true value in probability, the better is the method of boundary construction. Integrated nuclear density or DNA absorbance was the invariant selected.

Leukocyte features were derived in two ways: (1) by computations involving the partitioned density histogram, and (2) by processing the articulated gray-tone image. Although stable shape-oriented parameters, textural descriptors, and local detail can be obtained only by the second method, many relevant global density-oriented parameters can be obtained by the first, more economical method. Taking advantage of the internal organization of information in a white blood cell image, a wide variety of features can be calculated from the optical density histogram without referencing the scanned image a second time. These include simple intuitive features such as cell and nuclear area, mean cytoplasmic optical density, integrated cell and nuclear optical density (absorbance), and nuclear: cytoplasmic contrast. They also include more complex abstract features such as the kurtosis, skewness, and entropy of the optical density histogram.

Among the features which either cannot be obtained in this way at all, or else cannot be obtained with reliability or precision are cell and nuclear perimeter, mean diameters, asphericity or eccentricity, curvature-describing parameters, local texture descriptors, the histogram of local gradients, and the corresponding mean and variance. These require direct operations on the image itself.

The current set of 80 leukocyte parameters comprises 72 density-oriented parameters derived from the histogram, and 8 shape-oriented features, derived from spatial differentiation and contour tracing. These fall into five parameter families which have been assigned the following interpretive names: (1) contrast and texture, (2) nuclear-cytoplasmic comparisons, (3) concentration, (4) size and content, and (5) shape. These families were assembled on the basis of (i) operational definitions of the parameters and their interpretation in visual terms, (ii) the magnitudes of their within-cell type correlations, (iii) the degree of intersubstitutability in formal leukocyte discrimination by computer, and (iv) their independent participation in minimum-dimension best sets of discriminating parameters.

Estimation of the density-oriented descriptors from the histograms involved the computation of curve parameters such as moments through order four; modes and intermodal distances; antimodes and ranges; and certain algebraic combinations of these which represent nuclear:cytoplasmic comparisons. Parameters of the entire cell used portions of the histogram which are attributable to the cell. Parameters of nucleus and cytoplasm were computed from their corresponding subdivisions. Quantitative definition of shape required identification and labelling of boundaries and connected components in articulated images. In general, there was negligible observed difference in leukocyte feature values and in discrimination performance with minor changes in background reference level or in boundary thresholds, and the best features from the point of view of discrimination were perturbed the least.

A description of leukocytes by numeric characterization of their images leads naturally to vectorial representation and the decision-theoretic approach to discriminatory logic. A cell having measured values $x_1,\ldots,x_p$ for p features $f_1,\ldots,f_p$ taken in an arbitrary but fixed order was represented by the vector $X = (x_1,\ldots,x_p)$ in a p-dimensional space Ωp, in which each coordinate axis is associated with a specific feature. Since the coordinate values are measurements, X is a random vector of observations. The vector space Ωp spanned by all possible random vectors constituted the feature space or parameter space for the

discrimination problem. Further, for each of the g(=5) leukocyte types, a corresponding cell image category C_k was postulated. The g categories were presumed collectively to account for all classificatory contingencies, and to exclude shared membership or overlap (exhaustive and mutually exclusive).

The vectorial representation maps cells into points and cell categories into point sets, in a space of limited dimensionality. Different selections of variates clearly may lead to topologically diverse descriptions of the universe of discourse. From the point of view of computer-implemented discrimination, feature spaces in which the images of the various categories correspond to compact and well-separated clusters of points are desirable. Part of the motivation for developing feature selections algorithms is to generate such well-behaved configurations in feature spaces of low dimensionality. This most favorable disposition is seldom realized. At best, points cluster in more or less disparate regions of the space in a configuration suggestive of the underlying taxonomy and the clusters generated by different categories interdigitate. Perfect recognition is unobtainable in principal, but we can instead look for the "best" performance possible under the circumstances.

Toward this end, a measure of discriminatory achievement must be defined, and an optimal classification strategy relative to this measure must be determined. Statistical decision theory was used to partition a feature space or selected subspaces into g mutually exclusive and exhaustive regions ($R_1,\ldots,R_g$) corresponding to the g alternative leukocyte types, $C_1,\ldots,C_g$. Identification or recognition then consists in assigning a cell to the category C_k if, and only if, its vectorial representation X falls into the associated region R_k. The partition is deliberately constructed so that long term recognition performance is optimized in some sense. The optimal strategy is a function not only of a selected performance measure and the constraints of optimization but of the particular combination of probability distributions peculiar to the problem at hand and of the relative seriousness of unacceptable decisions as well. The introduction of probabalistic concepts acknowledges that there may be an inherent ambiguity or indeterminacy attached to the discrimination process which sometimes makes unique characterization relative to a given taxonomic system unattainable.

Desirable discriminatory performance was construed to mean consensus between the computer and an acknowledged cytological expert, with respect to the complete set of preconceived classificatory alternatives. This goal was interpreted mathematically as an optimization problem directed toward minimizing the adverse consequences of erroneous or divergent decisions, or more specif-

ically, toward minimizing the probabalistic expected cost of "misclassification." Bayesian decision strategies are suitable formalizations for this because by design they maximize the probability of concordance or else minimize the penalties of discordance (computer "error"). These decision strategies optimize (maximize or minimize) an objective which is a function of (1) prior probabilities, p_i and $p(X|i)$, and (2) a loss or cost matrix, $L = [L(D_j|C_i)]$, specific to the application and stipulated in advance.

The prior information consists of (1) the relative frequencies $P = (p_1,\ldots,p_g)$, where p_i is the prior probability that C_i is the "correct" category for a randomly selected cell, and (2) the relative frequencies $\{p(X|i): \quad i = 1,\ldots,g\}$ of all the possible observations within each category (i.e., $p(X|i)$ is the conditional density for X when it does in fact represent a cell belonging to C_i). The unconditional probability density $p(X)$ of the random vector X, independent of its origin, is the same as the joint density of the p parameters $x_1,\ldots,x_p$, and it is a linear combination of the preceding two probabilities. Bayes inversion formula relates these probabilities to $p(i|X)$, the conditional probability that the specific vector X represents a member of C_i, that is, to the posterior probability of C_i:

$$p(X)\ p(i|X) = p_i\ p(X|i) \text{ for all } X \text{ and } i = 1,\ldots,g,$$

where

$$p(X) = \sum_{i=1}^{g} p(X|i)\ p_i\ .$$

Element $L(D_j|C_i) = L(j|i)$ of the loss matrix expresses the penalty (+) or benefit (-) incurred by deciding D_j in favor of category C_j, when the subject actually belonged to category C_i.

Anticipated discriminatory achievement depends not only on the loss matrix L, which reflects the utilitarian importance attached to the various classification outcomes and is independent of any decision rules, but also on an event matrix $E[R]$, which gives the probabilities of the various classification possibilities under the yet to be selected decision rule R. Corresponding to the element $L(j|i)$ of L which stipulates the penalty or benefit associated with assigning a member of C_i to C_j is the element $E(j|i,R)$ of $E[R]$ which gives the probability of the same event. The event matrix is a function of the category regions $\{R_k\}_{k=1}^{g}$ induced in Ω_p by the decision rule R, as well as a function of the posterior probabilities $p(X|k)$:

$$E(j|i,R) = \int_{R_j} p(X|i)\, dX.$$

The expected risk under the decision rule R is defined as the long-term overall cost of classification

$$C[R] = \sum_{j=1}^{g} \sum_{i=1}^{g} L(j|i)\, E(j|i,R)\, p_i$$

$$= \sum_{j=1}^{g} \int_{R_j} \sum_{i=1}^{g} L(j|i)\, p_i\, p(X|i)\, dX .$$

In terms of conditional average classification costs for an individual decision on the vector X favoring C_k,

$$L_X(k) = \sum_{i=1}^{g} L(k|i)\, p(i|X) = \frac{1}{p(X)} \sum_{i=1}^{g} L(k|i)\, p_i\, p(X|i)$$

this overall cost is equal to

$$C[R] = \sum_{k=1}^{g} \int_{R_k} L_X(k)\, p(X)\, dX .$$

Note that C[R] is a function of the vector of prior probabilities, $P = (p_1,\ldots,p_g)$. By definition, the Bayesian optimal classification rule R* minimizes the (overall) expected cost of classification C[R] . This is ensured by favoring the category with the smallest conditional cost $L_x(k)$ for an individual decision about X. Equivalently, it is ensured by maximizing a discriminant $D_k(X)$ which is a strictly decreasing function of $L_X(k)$, for example, $D_k(X) = \ln (1 - L_X(k))$ over all k. Thus the optimal Bayes' strategy is given by the following rules R*:

R*: assign X to C_k where $D_k(X) \geq D_i(X)$ for $i = 1,\ldots,g$,

and the corresponding system of best regions R* is given by

$$R_k^* = \left\{ X \,\middle|\, D_k(X) = \max_{1 \leq i \leq g} D_i(X) \right\} .$$

The corresponding minimum cost C[R*] is obtained by using the best regions R_k^* in the formula for C[R], and represents the theoretically ideal or optimal performance.

For the special case of symmetric cost, algebraic simplification shows that the average loss is also the probability of misrecognition. Bayes' strategy favors the most likely category and

the best regions are

$$R_k^{**} = \{ X \mid p_k\, p(X \mid k) \geq p_j\, p(X \mid j),\ j = 1,\ldots,g \}\ .$$

This is the unconditional maximum likelihood decision rule. Ignoring or equating prior category probabilities ($p_i = p_j$, for all i,j), the Bayes rule becomes the well-known (conditional) maximum likelihood strategy:

$$R^{***}:\ \text{assign } X \text{ to } C_i \text{ if } \lambda(i,j) = p(X \mid i)/p(X \mid j) \geq 1.$$

It favors assignment of X to the most likely source, independent of category frequencies in the composite population, and maximizes the average success rate over the g groups, $1/g\ \Sigma_i E(i\ i,R)$. The discriminants reduce to

$$D_k(X) = \ln p(X \mid k).$$

The analytic form of the discriminants is determined by the specific probabalistic assumptions about the cell populations and the cost matrix. Linear and quadratic discriminants are technically optimal when the conditional densities $p(X \mid i)$ are unimodal, ellipsoidally symmetric multivariate densities, and the loss is symmetric (i.e., when a decision is always forced and the loss function assigns a standard penalty for erroneous computer decisions but none for correct decisions). The discriminants are monotonic functions - logarithms, for example - of the quantity

$$a_p \mid \Sigma_i \mid^{-\frac{1}{2}} f((X - U_i)^T \Sigma_i^{-1} (X - U_i))\ p_i$$

where a_p is a positive constant depending only on the dimensionality of the feature space, U_i is the multivariate population mean for C_i, Σ_i is the corresponding dispersion matrix, and f is a monotonically decreasing function of its argument. Multinormal probability densities fall into this category. Probability isopleths for such distributions are hyperquadrics in Ω_p. Linear discriminants and hyperplane decision boundaries are optimal for populations which differ in location U_i but not in dispersion Σ_i. Quadratic discriminants and hyperquadric decision boundaries are optimal when the dispersions are not identical. The discriminants for general multinormal discrimination are the family of functions

$$Q_i(X) = (-\tfrac{1}{2}\, p \ln 2\pi) + (\ln p_i - \tfrac{1}{2} \ln \mid \Sigma_i \mid\ ) - \tfrac{1}{2}(X-U_i)^T \Sigma_i^{-1} (X-U_i)$$

which specialize to linear functions in the homoscedastic (equal covariance) case. The insensitivity of linear discriminants to heterogeneous dispersion and moderate departures form normality is well documented and they perform impressively well even under sig-

nificant departures from the less demanding hypotheses which guarantee their optimality.

Our goal was to achieve machine discrimination of leukocytes that was as compatible as possible with our own discrimination, independent of cell type and diagnostic intent. Our own discriminatory ability is reflected in the decision-theoretic model when we specify and estimate the conditional probability distributions, $p(X|k)$, and the priors, p_k. By insisting on indifference of machine performance to cell type, we meant first, that all divergent decisions (computer "error" or "misclassification") are equally undesirable, and that all convergent decisions ("correct" classification) are equally meritorious; second, that compatibility of individual decisions was more important than long-term consensus for sequences of decisions; and third, that recognition performance should be uniform for all cell types. These subjective goals were best rendered by selecting a symmetric cost matrix and equating prior probabilities. This is tantamount to assuming the greatest uncertainty about the composition of the source population in the information-theoretic sense. The resultant maximum likelihood strategy is appropriate in the absence of any prior knowledge, or when all categories are to receive equal emphasis.

Were it not for the facts that the probabilities p_k and $p(X|k)$ are unknown, and that the penalties and benefits $L(D_i|C_i)$ are difficult to quantitate, discriminants could be constructed, constants could be evaluated, and automated Bayesian recognition would become a straightforward computational exercise. In reality, compromise with the strict definition of optimality is unavoidable. We have followed the practice of using an actual decision rule which approximates the ideal in a statistical sense, and retains the same hypothesized analytic form. Population parameters of discriminant constants are then estimated from prototypes for each category which have been selected and "correctly" labelled by an accepted authority. The actual rule is the sample analog of the Bayes rule, wherein the population means and common dispersion matrix U_j, U_k and $\sum$ have been replaced by their unbiased, maximum likelihood estimators, $\bar{U}_j$, $\bar{U}_k$, and $\bar{\sum}$. These estimates are consistent and approach the population values with increasing accuracy as sample size increases.

Thus the method of choice for leukocyte recognition was nonadaptive, supervised linear discrimination. The linear discriminants and polygonal decision regions were those which would have resulted from minimizing the composite error for equally likely multivariate normal (or ellipsoidally symmetric) populations with common dispersion under symmetric loss. In our application, category dispersions were unequal. Nevertheless, linear dis-

crimination was successful because the category means were so far apart in p-space.

The Mahalanobis distance can be generalized to several groups and tests of significance for composite discrimination can be based on the extended form. The statistic is

$$D^2_{pg} = \sum_{k=1}^{g} N_k(\bar{U}_k - \bar{U})^T \Sigma_{pg}^{-1} (\bar{U}_k - \bar{U})$$

where N_k sampl-s of C_k are available, Σ_{pg} is either exact or has been estimated on a large number of degrees of freedom, and $\bar{U}$ is the overall mean. D^2_{pg} can be used as a chi-square with $p(g - 1)$ degrees of freedom to test the composite hypothesis that the mean values based on p features are the same for all populations.

The pairwise tests of significance can be extended in yet another way to test the composite hypothesis $H_o : U_k = U_j$ under the assumption of common covariance, against the alternative $H_1 : U_k \neq U_j$. This extension proceeds from the interpretation of Fisher's discriminant as the maximum attainable ratio of between-to within-class variance for linear compounds of the original variables $x_1,\ldots,x_p$. The heart of this method is the analysis of scatter or dispersion into within-class and between class components, and the introduction of the generalized variance ratio, the ratio of the determinants of these matrices. The test statistic, the Wilkes-Bartlett lambda, varies inversely with the proportion of total dispersion attributable to category differences, and furnishes a formal compound test of equality for the g group centroids, under the hypothesis of common covariance. Lambda also plays a role in feature selection.

Since the number of features which we consider is limited only by our own ingenuity and patience, we are often confronted with more variables than we would like or can accommodate. It is therefore appropriate to ask (1) which, if any, variables in the current set can be discarded without impairing the discrimination significantly, and (2) can subsets of features surpass the diagnostic performance of the complete set. Feature selection refers to the systematic screening of the original set of variables for effective and computationally tractable combinations. This is a multiphasic endeavor, involving: 1) feature selection per se - inclusion of variables which facilitate discrimination; 2) feature reduction - exclusion of variables which are redundant or irrelevant, and hence possibly unwanted; 3) feature combination - formation of new variables with desirable mathematical and statistical properties from the original variables; and 4) feature ordering - serial arrangement of possibly independent or uncorrelated variables according to their individual discriminatory merit.

The impetus for utilizing multiple features is many-fold. First the theoretical Bayesian error rate cannot be affected adversely when the feature set is augmented although definite theoretical benefits have been demonstrated for only some special cases. If the additional variable is at all relevant and not totally redundant, the error rate must decrease. Similarly, between-group contrast, as measured by Mahalanobis distance, cannot decrease and most likely will increase. If these trends are rapid enough, improving discrimination by adding variates is reasonable to the extent that it remains practical. Finally, the contribution of features to discrimination may be indirect because of correlations. Retention of relevant but covariant features is seemingly inefficient, yet intuition suggests that redundancy might provide a corroborative effect. The outcome is determined by the relationship of interspecific to intraspecific variability and correlation. In individual features, a high ratio of interspecific to intraspecific variability is desirable. In pairs of features, a high intraspecific correlation with interspecific correlation of opposite sign is desirable. Relatively poor discrimination with a given set of features sometimes can be dramatically enhanced by the addition of suitably correlated features which by themselves are not diagnostic. However, correlation and not dimensionality is the primary factor operating. Uncritical addition of variables with reinforcing rather than contrasting patterns of intragroup and intergroup correlations can be self-defeating.

From the point of view of methodological consistency, the appropriate criterion for choosing among alternate sets of features is the system performance measure: the best set of features, relative to the totality available, is the one with the minimum expected cost of classification, or alternatively, the one with the minimum expected error rate. In other words, "best" feature subsets correspond to "best" optimal discrimination. Accordingly, the outcome depends on the behavior of the Bayes' risk and, in practice, the behavior of the actual risk, when the p-dimensional feature space is restricted to a p' ($<p$)-dimensional subspace.

The necessary computations for feature selection to comply with the optimality criteria are formidable. Since, in general, there is no assurance that the best set of p variates will have members in common with the best set of p' variates, the entire screening procedure must, in any case, be repeated when new parameters become available. Short of exhaustive search, there is no simple comprehensive optimal feature selection algorithm. The other extreme, random selection, seems too casual. A practical alternative is a limited, but systematic search which avoids the unwieldy complexity of exhaustive search, and incorporates a modified performance criterion which is responsive to the inherent resolving power of the variables.

The relationship of error rate to Mahalanobis distance in the pairwise normal case provided a rationale for an alternative figure of merit: maximization of the between-class to within-class variance ratio which is equal to the Mahalanobis distance in this case. The figure of merit which we adopted was its multivariate multi-category extension; more precisely, we used the total to within-class generalized variance ratio of Wilkes.

One way to reduce search time is to substitute stepwise, accumulative procedures which are exact in well known special cases, and generate a nest of maximally effective features in these instances. The essence of a sequential algorithm which we followed is to compute at each stage, a context dependent measure of effectiveness for each of the available but not yet selected variables, and retain that feature which gives the largest value of this measure in the presence of the subset of already selected features. Hopefully, by incorporating the most informative feature at each stage, an acceptable performance level will quickly be achieved and the selection process terminated in a low dimensional feature space.

The stepwise search derived sets of features, each extending the capability of its predecessor as much as possible relative to the figure of merit. The subset is developed by beginning with the single most powerful feature and iterating a two phase rule for introducing or eliminating features one at a time. The first or additive phase tries to improve the discrimination; the feature which leads to the greatest increase in total to within-class generalized variance ratio in the presence of the already selected features is added, provided that this fractional increase exceeds a stipulated level of statistical significance; otherwise, nothing is added. The second or subtractive phase tries to control redundancy; the feature which contributes least to the generalized total to within-class variance ratio in the presence of the remaining features is deleted, provided that the fractional decrease does not exceed a stipulated level of statistical significance; otherwise, nothing is deleted. The algorithm terminates when either the list of available features has been exhausted or tolerance levels preclude further changes. The computations at each step automatically compensate for discriminatory power already achieved by the previously selected features by the method of covariance adjustment.

Linear discriminants based on the current selection of variables were computed for each pair of cell types (neutrophils, lymphocytes, monocytes, eosinophils, and basophils) in each stage, and the confusion matrix giving the corresponding classification results was also derived. In this application, an initial set of 80 parameters was reduced to a set of four which achieved 100 per cent correct classification of leukocytes: conrast (coef-

ficient of variation of cell optical density), absorbance ratio or ratio of nuclear optical density to cell optical density, mean cell density (mean brightness), and nuclear absorbance.

Experience indicates that the resultant feature sets are near optimal. However, features selected in the opening steps may not always yield the lowest possible error rates for the corresponding dimensionality. This occurs when intergroup differences are unevenly distributed among the groups; the variance ratio, which is used to identify strong features, is sensitive to the larger inter-group contrast, and it biases the selection rule in favor of the variables which make the easiest pair-wise discriminations. Features which are accepted early in the development because of their discriminatory strength may later prove ineffective in combination with others, and will be discarded, not because they are weaker than subsequently selected features, but because they are redundant. The continual critical reappraisal of accepted features and rejection of useless ones is thus an important safeguard in the selection algorithm.

Canonical analysis, a bi-product of the feature selection algorithm, determines the linear transformation to the smallest subspace which accounts adequately for data variability, and assigns proportions of the variability to the selected dimensions. By definition, the new coordinate axes are in the direction of the greatest inter-class to intra-class dispersion, and the transformed features are uncorrelated and standardized (have unit variance). The ratio of total generalized variance to within-class generalized variance in the canonical space is partitioned so that the first axis in the new coordinate system has the largest quantum, the second axis has the next largest quantum, and so on. It was shown that canonical analysis in a four-space adequately accounted for the generalized variance ratio, and that on the basis of the coordinate axes' weights, four feature families could be identified.

Although most of the cumulative variability can generally be accounted for in a canonical space of lower dimensionality than the original space, all of the original variables are needed to compute the new coordinates. The canonical analysis is not a genuine method of dimensionality reduction: it is merely a guide to what could be achieved with linear discriminants and uncorrelated parameters. The ordered set of eigenvalues of the product of total and inverse within-class covariance matrices give an empirical indication of the number of linearly inde-

pendent factors operating. It can be shown that these eigenvalues are related to the Mahalanobis inter-group distance, and hence to the ideal error rate for the two category Gaussian discrimination problem.

The five white cell types were represented as well disbursed, compact clusters even in the space of the best three features. The intense unimodal clumping and the high degree of separability in the original feature space and especially in the canonical coordinate system of standardized and uncorrelated variables, make apparent the reasons for the success of linear functions, although technically they are not the optimal discriminants. These observations strongly suggest that the five cell types might be recovered from the raw data without the need for a learning set or for an analysis based on distributional assumptions, e.g., by a clustering technique.

A non-parametric cluster synthesis was therefore developed, using similarity or proximity in feature space rather than maximum likelihood as the basis for class membership (Paper VI). Unlike the statistical approach, neither the nature nor the number of categories was or need be prescribed. The notion of cluster is explicated in terms of quantitative statistical, psychological, and mathematical notions such as resemblance, proximity, tendency, connectivity, and significance. Pattern articulation procedes de novo on the basis of heuristics embodied in the algorithm rather than on the basis of externally supplied information. Thus a taxonomy responsive to structure detected in the raw data is generated in preference to a superimposed, preconceived codification.

The clustering algorithm accepts object parameters or similarity relations, and generates groups by (1) examining a dynamic dissimilarity function defined for all pairs of objects and clusters and constituting an objective measure, (2) applying a merger strategy in accordance with iterative optimization of the dynamic objective function, and (3) evaluating a dynamic objective measure which can be thresholded to provide a halting criterion. Program options permit variation of data normalization, measures of similarity, and clustering strategies.

There are three principal phases in cluster synthesis. In the first or preprocessing phase, parameters of the candidates can be transformed by one or more mappings which fall into the category of linear transformations. Among the possibilities are pure translation, pure rotation, feature selection, suppression or accenuation. Variables typically are centered about the universal centroid, and scaled by their standard deviations.

The second phase consists of computing a measure of dissimilarity or distance for all candidate pairs,using one of a variety of metrics or similarity coefficients. These include Euclidean and other Minkowski metrics, including the limiting case, various correlation and matching coefficients, and a measure of non-random association. Alternatively, the synthesis may procede directly from similarities rather than object parameters.

In the final phase, clusters are synthesized. Initially, each object is a singleton cluster. The "most similar" or "closest" clusters are fused to form a new cluster which replaces them. Dissimilarities or distances between revised clusters and unperturbed clusters are computed in accordance with the elected clustering strategy, and the entire process is iterated until the supply of candidates for merger is exhausted. The elements of the dissimilarity matrix thus constitute an objective function which is minimized at each iteration and drives the formation of clusters. This progressive modification of the dissimilarity matrix involves one of six variants of a single transformation which can direct the mergers so as to emphasize either compactness, connectedness, or central tendency of the resultant clusters:

$$\delta^2_{tr} = \alpha_p \delta^2_{tp} + \alpha_q \delta^2_{tq} + \beta_{pq} \delta^2_{pq} + \gamma_{tpq} \left| \delta^2_{tp} - \delta^2_{tq} \right|$$

where appropriate values are assigned to $\alpha_p, \alpha_q, \beta_{pq}$ and γ_{tpq}. The algorithm can thus efficiently accommodate a broad range of cluster shapes and topologies within a parsimonious program structure.

Since cluster formation involves successive merger of preexisting clusters and isolated objects until all objects have been subsumed by the universal cluster, and since a membership assignment, once made, cannot be revoked, a hierarchy or tree of nested clusters encompassing the entire sample is generated. The root corresponds to the conjoint partition, branch tips correspond to the disjoint partition, and nodes represent intermediate arrangements. The tree is equivalent to a nest of increasingly coarser partitions of the universe of discourse into equivalence classes.

The status of the clustering is summarized by successive values of an objective measure which is a generalization of the multivariate within-cluster to total variance ratio, computed either in the space of transformed variates or else in Euclidean space. Within-cluster "variance" increases continually as clusters are

fused. Since the total "variance" or sum of squared dissimilarities is fixed, there is inevitable reduction in between-cluster "variance." The objective measure is therefore bounded and ranges from a value of 0 for the disjoint partition (no internal heterogeneity) to a value of 1 for the conjoint partition (no external heterogeneity). Changes in the objective measure are computed at the same time that inter-cluster distances are revised.

Both objective function and objective measure are monotonically increasing functions of the number of clustering events. In general, clusters which emerge because they are currently "closest" or "most alike" do not correspond to the smallest possible decrement in between-cluster "variance." However, for one of the merger strategies dissimilarity between two clusters is measured by the incremental internal variance of the hypothetical conglomerate; minimizing the objective function at each iteration in this case coincides with minimizing the objective measure, and cluster formation tends to preserve within-cluster homogeneity or internal structure.

Values of the objective measure are useful for evaluating the significance of clusters. The sequence of values of the objective measure at the completion of each iteration corresponds one-to-one to the sequence of clustering events by definition, and these values can be used for determining a halting criterion. The cutoff threshold is at the users' discretion. In general, the level which seems to correspond to psychological significance is a function of the underlying probability structure of the data, as well as of the cluster merger strategy. The selection of a cutoff threshold is analogous to the selection of significance levels in statistical hypothesis testing. In our diversified experience, clusters existing at the 0.01 level of the objective measure were persistently in harmony with independent human judgments. For multivariate normal populations with common covariance which are analyzed in Euclidean space, the objective measure equals Wilkes' lambda statistic, and it leads to an exact statistical test of significance of global discriminatory achievement.

The application of clustering to leukocyte images had two purposes: first, the de novo construction of a cytotaxonomy based on machine rather than human perception and cognition; second, validation of the traditional leukocyte taxonomy which formed the basis of the statistical approach by prescribing the learning sets. Using the algorithm as described above, recovery of the five white cell types was completely effective at the 0.01 level of the objective measure. The five categories were stable over a broad range of normalization procedures, metrics and cluster merger strategies.

For Minkowski spaces with exponents 1.0, 1.5 and 2.0 classification errors on a sample of approximately 100 cells were only 1%, 1% and 4%, respectively. The nodal positions marking merger of the main cell types varied with the metrics but the categories maintained a regular pattern of relationship.

In this way, taxonomic consistency was established between the unverbalized qualitative discriminatory criteria used by the morphologist, and the feature space based on quantitative computer-generated, tonally oriented leukocyte features. By this we mean that the morphologist's classification of white blood cells and the numerical taxonomy agreed in extension, although they were based on fundamentally different criteria. This, we felt, was a fundamental test for not only the clustering algorithm but also for the approach to characterizing leukocytes.

In summary, the objective of our leukocyte study was to distinguish the five types of white cells found in normal human peripheral blood on the basis of computer processing of digitized microimages. The research indeed indicated that machine perception can contribute materially to cytodiagnosis, not merely in performing lesser tasks, but in stimulating fundamentally new insights in cytomorphology and cytophotometry.

The densitometrically oriented approach to the analysis of cell images which we introduced emphasized: (1) finely discriminated gray-scale information taken at high optical resolution, (2) the use of optical density frequency distributions of images to perform the preliminary spatial organization of quantitative optical information, (3) the extraction and evaluation of large collections of potential image-characterizing parameters within a decision theoretic framework, and (4) the algorithmic construction of a taxonomy for mature leukocytes.

Experimental results have validated the density oriented-approach to the analysis of cell images, and have led us to the following position:

1. The algorithms for image articulation or feature extraction generate image parameters that are responsive to taxonomic differences among leukocyte types, yet relatively insensitive to within-type variability and to secondary factors attendant in specimen preparation and electro-optical scanning.
2. The performance criterion of maximizing the compatibility of man-made and machine-made leukocyte discriminations can be formulated in decision-theoretic terms, and leads to efficient implementation of discriminatory logic in the form of linear discriminant functions, operating in a multidimensional vector space of image representations.

3. The set of parameters is sufficiently rich to characterize the five leukocyte types: lymphocytes, monocytes, neutrophils, eosinophils, and basophils. Many subsets of as few as four parameters yield effective machine discrimination.
4. Statistical and quasistatistical or heuristic methods exist for feature selection, that is, for obtaining strong minimal subsets of characterizing parameters and for evaluating their absolute and comparative discriminatory effectiveness. This systematic assemblage is superior to parameter selection on the basis of intuition and random search, and is more efficient than exhaustive search and comparison. Furthermore, it provides insight into the variance-covariance structure of good feature sets.
5. Multivariate linear discrimination is approximately optimal for leukocyte discrimination with respect to the dual criterion of minimizing man-machine divergence in discrimination while minimizing the requisite dimensionality of the image-representing space.
6. The parameters can be given plausible morphologic and densitometric interpretations, and hence correlated with cell descriptions. The near-optimal subsets of cell parameters give insight into the salient global visual differences and similarities among blood cell images.
7. Densitometric information, as contrasted with morphological information, is totally adequate for cell identification, and indeed was somewhat more powerful.
8. The ensemble of algorithms for cell articulation, characterization, and identification appears to be a robust solution to the problem of machine discrimination of normal mature leukocytes.

The success of the linear discrimintory logic is attributable in part to the appropriateness of the methods of image articulation to the subject matter at hand, in part to the plethora of computer-generated features, and in part to the robustness of the decision-theoretic and feature selection techniques.

The methodology developed for leukocyte discrimination is also applicable to pattern recognition in other cell systems (Paper V). It was used to classify cells from a routine fine needle breast aspiration biopsy smear into four categories - lymphocytes, macrophages, stroma cells and cancer cells - using five parameters with virtually no error. The five features used in linear discriminant analysis were entropy, kurtosis and skewness of the optical density histogram, and average chord length and area of each cell.

4. AN APPLICATION TO CYTOGENETICS

Somatic cells are observed in metaphase, the middle stage of mitosis or cell division, since this is when the chromosomes are condensed and most visible. In addition, the nuclear membrane has disappeared, so that the chromosomes have a larger area in which to orient themselves. The division process is arrested at this stage, and the cell samples are fixed, mounted and stained for microscopic investigation. The 46 chromosomes of the normal human cell, which are oriented in a seemingly random fashion, can be reduced to 23 chromosome pairs (22 pairs plus two unmatched sex chromosomes in the male). The pairs can be arranged in an arbitrary but standardized fashion, called a karyotype, by determining total chromosome length and location of the centromere (the point of apparent contact of the two chromatids which make up each metaphase chromosome or the point of constriction of the two sister chromatids). Chromosome pairs bear a numeric designation and sets of paired chromosomes which resemble each other are further designated by seven alphabetic group labels (plus the sex chromosomes). Additional clues used to aid in the identification of particular chromosomes include the presence of "satellites", small masses of chromosomal material suspended above the short arms of D and G group chromosomes but connected to the main body by a thin thread; and secondary constrictions, or thinned out sections of the chromosome arms, appearing consistently in various locations on certain chromosomes.

Usually each normal mammalian metaphase chromosome contains two identical chromatids which are either joined together or in close opposition at the centromere (major constriction). The chromatids have an underlying multistranded, multispiraled structure. There is no suggestion of a sharp or well-defined external chromatid membrane, and the combined structure is better modelled by a ragged test tube brush than by a smooth sausage. There is a primary constriction or narrowing of the chromatids associated with but not necessarily identical to the position of the centromere. Secondary constrictions may occur elsewhere along the structure, and satellites or terminal appendages may be present. Abnormalities include fragmentation, gaps, additional or missing material in one or both chromatids, and extra centromeres. The resultant image is a fuzzy, somewhat irregular, often asymmetrical pair of interacting curvilinear structures. Sister chromatids can usually be resolved, but they are often in such close proximity that they blend together to a considerable extent.

The purpose of processing the somatic cell chromosomes as described is to facilitate the detection of deviations from the normal pattern. There are characteristically 46 chromosomes in man, and a consistent deviation from this number is of biomedical

importance. In addition, structural changes such as breakage of chromosomes with (in some cases) translocation of material, result in chromosomes which no longer have the appearance of the original ones. These structural abnormalities may comprise loss of material with fragments possibly present elsewhere in the cell, and abnormal rejoining of broken ends, either when both ends of one chromosome were broken and rejoined to form a ring, or when two chromosomes were broken and then joined so that a chromosome results with two centromeres instead of one, a so-called dicentric chromosome. Abnormalities in cell division may lead to endreduplicated chromosomes, with extra associated chromosome duplicates of the usual ones, and quadriradial structures in which the chromatids of two chromosomes are closely associated.

Some congenital abnormalities manifest themselves in chromosome karyotypes with a characteristic clinical abnormal appearance. Radiation, chemicals, viruses and other environmental changes sometimes produce chromosome aberrations.

Paper VII describes strategies for scene segmentation and image articulation in detail.

During image processing, the digital image of a chromosome undergoes a variety of transformations. The graphical presentation reduces each chromatid to a single line. The underlying architecture of the chromosome is efficiently displayed, including the location of the centromere and the length of the chromatid arms, but various aspects of shape and the gray information in the image are discarded. A black-and-white silouette or outline drawing of the chromosome distills the shape information by means of an idealized external boundary. It is achieved in digital images by selecting an appropriate brightness threshold and generating a one-bit image.

In the photometric version of the digital chromosome image, optical density, transmittance, or some other function of light intensity is accurately translated into grayness. Optical density in the image is proportional to the local concentration of light-absorbing material in the object, and the sum of optical densities of the chromosomal image becomes a measure of the stain content of the chromosome. Chromosome boundaries must now be extended far enough beyond the silhouette of the chromatid to ensure the inclusion of all grayness attributable to that chromatid, and attention must be paid to apportioning the optical densities between adjacent or overlapping interacting objects.

Having obtained the digital image of the metaphase, the next steps are to isolate the individual chromosomes and to articulate the subordinate chromosome parts, the arms, centromere,

secondary constrictions and satellites. A variety of measurements can be made on the parts, and these extracted parameters form the basis of image characterization and karyotype analysis. The generation of a karyotype from chromosome measurements may proceed in a variety of ways. Pairing of homologues can be emphasized at the outset, as we did, or major groups can be assembled before individual chromosomes are otherwise committed. One procedure is to use sequential logic, identifying the most easily discerned members first, and proceeding step by step toward a complete and coherent analysis of the residual. There is also the opportunity to use the simultaneous logic of multivariate statistical analysis or clustering techniques (Paper IV).

Our approach to chromosome analysis emphasized photometry and the extensive use of tonal information in the image. Photometry is stressed because biological considerations suggest that measurements of DNA content will be useful, relatively stable descriptors of the chromosome. DNA is the primary carrier of genetic information, and current understanding of the chemical expression of DNA coding and of DNA replication leads to the hypothesis that the amount of DNA per ploidy set is a fixed characteristic of each genetic type. Cytophotometric measurements in the ultraviolet on the natural absorbance of DNA, and in the visible on histochemically stained DNA, confirm this constancy within the limits of the methods. Unlike parameters derived from length and area, parameters based on content have the potential virtue of being insensitive to chromosomal condensation and preparative flattenning.

The indication of the presence of a potential chromosome in an image is the finding of a high grayness level or a steep gradient of grayness. The object can then be delineated by thresholding or gradient tracking procedures. The thresholding method is based on the existence of a single gray level that will effectively divide a metaphase into discrete intact chromosomes. Too low a threshold will cause adjacent chromosomes to coalesce and too high a threshold will make individual chromosomes fragment into separate parts.

One approach to finding appropriate threshold levels is to use the method developed for white blood cell boundaries. When structural compartments of images differ significantly in grayness, the boundaries between compartments consist of scan elements with intermediate and infrequently occurring values of optical density. Chromosome images involve two compartments, and their optical density histograms have accordingly only two peaks and one trough. The optical density at the relative minimum of the histogram is a good threshold for generating outline drawings

of chromosomes. This threshold is closely correlated with the optimal threshold for chromosome isolation as derived from fragment number and coalescents number versus optical density.

Another method of finding the threshold is to find the mean optical density of the background scan elements. Since empty fields scanned by CYDAC give symmetrical frequency distributions of background optical density, the background of a metaphase can be extrapolated from the left-hand portion of the background peak of the histogram of optical density. Background can, therefore, be subtracted from such compound frequency distributions, and we have found that the mean optical density of the remaining points provides a good estimate of the optimal threshold level for chromosome isolation.

For locating centromeres in CYDAC images a method was based on the narrowing of the chromosome outline at the primary constriction or centromere. Experimental programs have also generated skeletons of chromosomes and have located the centromeric region in the process. Both of these methods begin with a one-bit bounded chromosome image. Photometrically oriented methods were also developed. Optical density profiles taken perpendicularly to the main axis of a large chromosome show some contrasting features of the centromeric and noncentromeric regions. Away from the centromere the chromatids separate enough to be partially or completely resolved in the digital image, appearing as twin peaks in the profiles. Through the centromere, the profile shows a single peak. The profiles suggest that the height of the peak is less at the centromere than elsewhere in the chromosome. The method can be greatly reinforced by summing the optical density of all points in each profile normal to the axis, and by combining together the totals for both chromatids.

In Paper VIII, the potential value of DNA content as a parameter for karyotype analysis is presented, and a method is developed based on integrated optical density measurements of gallocyanin-chrome alum stained human chromosomes. The optical information was obtained automatically in digital form by the CYDAC digital scanning cytophotometer. The digital images were analyzed by computer after being given the location of the centromeres. Groups of chromosomes, corresponding to the Denver classification, were correctly identified. This partly automated computer analysis of the data provided the relative stain content of individual chromosomes and chromosome arms. Experience with three metaphases from the same human female indicates stable clustering of the results by the use of two parameters - the integrated optical densities of large arm to total - with sufficient discrimination among clusters to identify eleven discrete chromosome

classes. It would appear from the values obtained that gallocyanin-chrome alum stain content of individual chromosome arms correlates closely with length and ultraviolet absorbance values given in the literature. An extensive analysis of the error involved and their causes is presented. The integrated optical densities in CYDAC images gave an average replication error of 4.7 per cent for the stain content of gallocyanin-chrome alum treated human chromosomes, and an average error of 5.7 per cent for the centromeric index (long arm/total optical density).

Paper IX explores the possibilities of a one-dimensional karyotype analysis in which the position of the centromere plays no part. Ordering the data in terms of integrated optical density consistently places each chromosome within a few positions of its visually determined Denver classification.

The computer surveys the optical density array and determines the frequency of occurrence of each CYDAC value. As previously described, in such frequency distributions, the principal peak is due largely to background points. The mode of this peak provides an estimate of the global background level. A second peak is attributable to the chromosomes, and the minimum between the two peaks corresponds to the steep gradient of grayness on the periphery of the chromosomes. A threshold is set at the CYDAC value corresponding to this minimum, and points with CYDAC values above threshold are designated as core points.

Computer printout of the optical density array is prepared with slashes overprinted at the core points. This reconstituted image is examined by eye and edited in the following manner: (1) Clustered core points that are attributable to interphase nuclei or dirt particles are marked so that they will not be treated as chromosomes in the subsequent analysis. This is done automatically for any clusters that touch the edge of the scan or have less than four core elements. (2) The coalescence of adjacent chromosome cores is undone by marking the bridging core points as non-core points. (3) Fragmented chromosome cores are repaired by marking the missing elements as core points. The change of core designation in these operations does not affect the optical density of the point. The computer scans the edited image array and identifies each chromosome or non-chromosome core as an isolated body of connected points. A shell that is four scan elements thick is constructed around each core. Points on the shells associated with two or more neighboring cores are assigned to the core that is closest. Points that are neither core nor shell are designated background.

Each chromosome core and its associated shell is framed within a

tightly fitting vertically oriented rectangle of scan elements. Within such a rectangle, shell and core elements belonging to other cores are ignored during the subsequent processing. The mean background is computed from the sum of CYDAC values of background points within the rectangle. Mean background is subtracted from the CYDAC values of each element in the core and the shell, and the sum of these corrected CYDAC values is considered the integrated optical density of the chromosome. The mean values are arranged in order of decreasing magnitude. The 46 mean values are summed to give the integrated optical density of the metaphase, and the mean value for each chromosome is expressed as a percentage of the metaphase total. Finally, the chromosomes are given a photometric karyotype assignment. One of the bases of the Denver classification is the ordering of human chromosomes according to decreasing length. A photometric karyotype assignment was therefore based on a one-to-one correspondence between the sequence of decreasing optical density values and the Denver convention.

Measurements on seven human metaphases showed that:
(1) the replication error for the integrated optical density of an average-sized chromosome is 2.9 percent;
(2) the stoichiometry between gallocyanin-chrome alum and chromosomal DNA is affected by chromosomal compactification;
(3) the gross effects of compaction disappear when the data are normalized by metaphase stain content;
(4) on the average, 35 of the 46 chromosomes are properly placed in the Denver classification by a simple ordering of the chromosomal stain contents as a decreasing sequence;
(5) the two homologous number 1 chromosomes of this normal human female consistently differ by 11 per cent in stain content and by 3 per cent in length, and
(6) less striking but significant variability is found within other chromosome groups and may involve differences in DNA content or in compaction between the homologues or among members of the group.

5. APPLICATIONS TO HISTOLOGY

A massive effort has been expended on the application of computer-based biomedical pattern recognition and image processing to microscopic images of human cells and chromosomes - to cytology. However, virtually no application to histology, the study of tissues, has been made, despite the fact that histology is taken as ground truth in pathology. This is due in part to the difficulty of the subject matter. This is also due to the large computer storage demands and small cycle time demands needed in order to accommodate the necessarily large digital images, demands that

were met only recently. Paper X presents some initial progress in processing microscope images of tumors of the epithelium of human urinary bladder.

At the present time, the diagnosis - and consequently the prognosis - of any given case of bladder cancer is dependent upon the histologic features of the disease as interpreted by a pathologist. Despite the fact that the clinical impression of an experienced urologist is often of great accuracy, and despite the fact that the cytologic diagnosis of an experienced cytopathologist is also highly accurate, the clinician almost never undertakes definitive therapy without a histopathologic diagnosis establishing the malignant nature of the lesion.

The trained microscopist mentally catelogues the size, shape, color and texture of tissue constituents with the purpose of recognizing pathologies. Tissues are graded according to the overall pattern of constituent cells and their nuclei. The diagnosis of an anaplastic malignant epithelial tumor of the bladder poses no problem for the pathologist. The histologic characteristics of this lesion can be readily distinguished from the histologic features of normal epithelium, or even of epithelium from an inflamed bladder. The major diagnostic challenge arises in the case of papillary proliferative lesions in which the individual cells show only minimal deviation from the pattern considered to be normal for the transitional epithelium lining the bladder.

Many pathologists who have studied this problem believe that there are some characteristics of nuclei or of cells in such papillary proliferative lesions which are characteristic for either the benign or the malignant state. Other pathologists find the border line between nuclear and cytoplasmic features in benign versus malignant epithelium to be rather vague. Some who find this borderline to be indistinct are of the opinion that perhaps the abnormality in neoplastic epithelium of the well differentiated variety lies not so much in the presence of truly abnormal cells as in the distribution pattern of cellular and nuclear features in the papillary proliferative lesion. That is, in normal or inflamed bladder epithelium, the size, shape, chromatin distribution patterns, and other features of cell nuclei are generally distributed over the usual bell-shaped curve. In a low grade papillary neoplasm, qualitative histologic studies suggest that nuclear size, for example, of cells in the papillary lesion might be greater than the mean size of normal epithelial nuclei, but that there will be more large nuclei and fewer small ones than in the aggregation of normal cells. Thus, instead of the usual bell-shaped distribution pattern of nuclear sizes seen in normal epithelium, the mean size of nuclei in papil-

lary proliferative lesions will be skewed to the right. Quantitative characterization may sharpen our ability to discriminate these qualitatively obscure or uncertain distinctions in bladder epithelium.

Of particular interest is a study of nuclear size, shape, optical density or brightness, chromatin distribution patterns and other textural features which lend themselves to quantitative assessment.

Computer programs (Papers XI and XII) were written for the following specific tasks: (1) data acquisiton, (2) data representation and organization, (3) data display, (4) data analysis, and (5) data interpretation. Data acquisition programs handle the conversion of scanned material into a library of tissue sections and cells. Data display occurred within the framework of a conversational programming system designed especially for handling gray scale pictures (PEEP/DECIDE/GRAPH). The system allows the user to display a gray scale image on a display scope and then transform or modify the image interactively. The line printer and incremental plotter were also used for presenting scatter diagrams, graphs, statistical outputs and digests, and publication quality 32-graytone reconstructions of images and their transforms.

Data analysis applied image processing and pattern recognition to cell and tissue images directly, and to multivariate frequency distributions of their features. Image processing utilizes techniques for: (1) enhancement, (2) global feature detection, (3) picture segmentation into objects of interest, (4) local feature extraction, (5) object characterization and (6) scene description. Pattern recognition utilizes techniques for: (1) feature selection and ordering (parsimonious characterization of objects for the purpose of forecasting), (2) discrimination/classification using discriminant analysis and supervised learning, (3) discrimination/classification using cluster synthesis and unsupervised learning, and (4) construction of a taxonomy for bladder epithelium. The taxonomy will be based on numerical "aberrations" arising from morphological or structural "aberrations" in the cell and tissue images which reflect functional "aberrations."

For experiments on the quantitative characterization of the epithelium of the urinary bladder, epithelial tissues and papillary growths embedded in parafin were sliced 4.0 micra thick mounted on glass microscope slides and stained with hematoxylin to reveal cellular and subcellular morphology and nucleic acid composition. The specimens were scanned at the Jet Propulsion Laboratory's ALMS-2 vibrating mirror scanning cytophotometer. Intersample distance and spot size were both 0.5 micron. Scanning was performed in both

white light and at 530 nm to enhance nuclear-cytoplasmic contrast and at 630 x magnification. Final images contained approximately 50 to 100 cells and each measured 1024 x 1024 pixels. Analysis of the images was made using the PEEP/DECIDE/GRAPH programming system (Papers XI and XII).

Scene analysis was performed using two methodologies. The first involved global and local thresholding of the image, the second, computation of spatial gradients of optical density. Neither technique is universal when applied to clinical preparations, but thresholding is more economical and is the method of choice.

Features or state variables of interest for epithelial cells include: size, brightness (density), absorbance (integrated density), spectrum or chromaticity, (chemical) composition, texture, topography, morphology (shape or structure), conspicuity, symmetry/asymmetry, homogeneity/heterogeneity, index of maturity, and index of malignancy. Features or state variables of interest for epithelial tissues include: spatial distribution of cells and nuclei - orientation; local population density and population density gradient; frequency distribution of cellular and nuclear features; spatial distribution of features; gradient fields of feature distributions; tissue organization - order/disorder transition; homogeneity/heterogeneity;appearance of margins; index of differentiation; index of malignancy; and index of prognosis. Many of these features have been embodied in quantitative explications in PEEP/DECIDE/GRAPH.

A vector space representation was given to individual nuclei in two tissue sections using the quantitative explications $x_1,\ldots,x_p$ of features generated by PEEP/DECIDE/GRAPH as coordinates. In each of a sequence of subspaces, characteristic functions or discriminants were derived to differentiate nuclei according to tissue specimen of origin. The discriminant for each tissue was a quadratic or linear function of the coordinate features:

$$Q_i(x) = -\tfrac{1}{2}p\ln 2\pi + (\ln p_i - \tfrac{1}{2}\ln|\Sigma_i|) - \tfrac{1}{2}(x - U_i)^T\Sigma_i^{-1}(x - U_i)$$

where p is the dimensionality of the subspace, p_i is the prior probability, U_i is the mean feature vector and Σ_i is the covariance matrix for the i^{th} tissue section. The functions are optimal in the Bayesian sense of minimum expected error rate when the categories have posterior probability distributions which are unimodal,ellipsoidally symmetric multivariate densities and the loss function for incorrect classifications is symmetric. An "unknown" nucleus $\underline{x} = (x_1,\ldots,x_p)^T$ is classified as belonging to the tissue corresponding to the larger value of the discriminants.

The "best" subspace for discrimination was selected iteratively

using as a criterion the total to within class generalized variance ratio, Wilkes' lambda, which is a function of error rate and of Mahalanobis distance in the pairwise homoscedastic case. The subset is developed stepwise by beginning with the single most powerful feature and iterating a two-phase rule for introducing or eliminating features one at a time. The first or agglomerative phase tries to improve the discrimination: the feature which leads to the greatest increase in total to within-class generalized variance ratio in the presence of the already selected features is added, provided that this fractional increase exceeds a stipulated level of statistical significance; otherwise, nothing is added. The second or subtractive phase tries to control redundancy: the feature which contributes least to the generalized total to within-class variance ratio in the presence of the already selected features is deleted, provided that the fractional decrease does not exceed a stipulated level of statistical significance; otherwise, nothing is deleted. The algorithm terminates when either the list of available features has been exhausted or tolerance levels preclude further changes. The computations at each step automatically compensate for discriminatory power already achieved by the previously selected features by the method of covariance adjustment (Paper IV).

Linear and quadratic discriminants based on the current selection of features generated by PEEP/DECIDE/GRAPH were computed for two tissue sections of differing appearance, and the confusion matrix giving the corresponding correct and incorrect classification results was also derived. The confusion matrix plays no formal role in the feature selection algorithm. It provides an independent measure of its efficacy. Best discrimination as evidenced by the confusion matrix was achieved in a minimal space of dimension three. The three discriminating features were: nuclear absorbance, density and density variance. An average rate of 89% correct classification was achieved.

As a companion to this supervised learning, an unsupervised classification experiment was performed on the same two tissue images. The method employed a nonparametric, hierarchical cluster synthesis in Euclidean two-space based on similarity rather than maximum likelihood as a basis for class membership (Paper VI). The algorithm utilizes recursive computations to develop a hierarchy of nested clusters. The major components of the program are: (1) a (dis)similarity function, (2) a grouping or cluster merger strategy, based on optimizing a dynamic objective function identical with inter-cluster "distance", and (3) a halting criterion, based on evaluating a dynamic objective measure, the within- to total-class generalized variance ratio. The principal output is a tree, scaled by the objective measure, showing the

nest of cluster mergers.

For the two tissues, the features used were nuclear absorbance and density variance. Initially, in a Euclidean feature space, the distance between nucleus-representing points was computed and taken to be a dissimilarity measure. The numeric features were made comparable by the device of scaling each by its standard deviation over all nuclei. The cluster merger strategy was to coalesce "nearest" neighbors, where a definition of inter-cluster distance is supplied by the program and emphasized the formation of compact clusters. At each iteration, the inter-cluster "distances" are updated. The cluster tree for nuclei in the two tissues shows a correct classification rate of 77%, again indicating that nuclei from one tissue have a greater resemblance to each other than to nuclei from the other tissue.

Another experiment was performed in order to determine the relative orientation of nuclei within a tissue section and to compare the frequency distribution function of relative orientation for different tissue sections. Toward this purpose, the boundaries of nuclei were determined from thresholded images. Each boundary was expressed as a complex function $Z(s) = X(s) + iY(S)$ of arc length s from an arbitrary origin. Being a periodic function of period σ, where σ is the length of the boundary, $Z(s)$ may be expanded in a doubly infinite complex-valued Fourier series.

$$Z(s) = \sum_{n=-\infty}^{\infty} F_n \exp(2\pi isn/\sigma)$$

with coefficients

$$F_n = \frac{1}{\sigma} \int_0^{\sigma} Z(s) \exp(-2\pi ins/\sigma) \, ds \, .$$

Passing to the discretization,

$$F_n \quad \frac{1}{\sigma_o} \sum_{j=0}^{n} Z(s_j) \exp(-2\pi ins_j/\sigma_o) \Delta s_j$$

where s_j is the j^{th} sample point on the eight-connected boundary consisting of N points of total length $\sigma_o = s_n$ and Δs_j is the j^{th} step length equal to 1 or $\sqrt{2}$. F_o is merely the center of gravity of the boundary and without loss of generality is set equal to $0 + i0$.

By taking just the terms

$$Z(s) = F_1 \exp(2\pi is/\sigma) + F_{-1} \exp(-2\pi is/\sigma)$$

we obtain the best trigonometric elliptical approximation with semi-major axis of length $|F_1| + |F_{-1}|$, semi-minor axis $||F_1| - |F_{-1}||$ and at an angle $\frac{1}{2} \arg F_1 F_{-1}$ with the coordinate axes.

The distributions of angles of inclination of the major axes for the two tissue specimens were plotted relative to an arbitrary vertical angle taken to be the respective modes. Specimens of normal epithelium and low grade tumors tend toward a more parallel alignment of nuclei than do specimens of high grade tumors. The higher the grade of the tumor, the greater is the disruption of normal regular tissue architecture and the more toward random is the distribution of angles. The two tissue samples show intermediate distributions which are nevertheless different. These results combined with the results of the discriminant and clustering analyses would imply differential grading of the two tumors.

6. SPECIALIZED SOFTWARE

Pictorial data forms an integral part of many patient records. Such data takes the form of radiographs, blood smears, cervical smears, sonograms, chromosome karyotypes, etc. It is in the long-range interest of biomedicine to encourage the development of programming systems and languages oriented toward the scientist-user who engages in the analysis (on-going theory) of such pictorial data (on-going observation). PEEP, a collection of picture editing and evaluation procedures common to many pictorial data processing problems, is such a conversational, interactive yet rapidly expandable programming system. It deals with digitized images with respect to preprocessing, global scene analysis, local scene analysis, feature and relationship extraction, feature selection, and classification by supervised and by unsupervised learning. PEEP is habitable in scope, civilized in use, and rapidly expandable and extensible. It comprises three parts, PEEP proper, DECIDE and GRAPH, sub-systems for picture processing per se, medical decision-making and two- and three- dimensional computer graphics. It is meeting the demands of several picture processing applications to white blood cells, tissue characterization, aspiration biopsies and computerized tomography.

The approach we adopted involved providing the computer with (1) a knowledge of relevant technical English, albeit a very limited knowledge in the form of grammars, (2) programs for scanning, parsing, generating and interpreting with respect to this linguistic knowledge, (3) a universe of discourse for the computer to converse about and deal with - structured data bases and (4)

means to access a data base as needed, and to associate appropriate organized data from the universe with close-to-English processing requirements.

The permissiveness of all of English is beyond the present state-of-the-art for any practical man-computer-data interface. For a limited subject matter, however, a proper subset of the utterances of performance as contrasted with the sentences of competence (i.e., sentences that not only are legitimate in discourse, but are likely to be used) might well furnish the computer with a viable very high level artificial command language, Such an artificial scientific language should be habitable in scope, civilized in use, and rapidly expandable and extensible.

Habitability entails the unlikelihood of straying over language boundaries into unallowed sentences. In other words, however limited the computer's scientific sub-language may be, the user is able to express himself readily within it while conducting scientific investigations with a given data base. Civility precludes impairment of scientific investigations within the subject universe because of the language's artificiality or unnaturalness and dearth of desirable data structures and commands. In other words, the user will not be intolerably confined or encumbered by the sub-language's paucity. Expandability and extensibility are keys to both habitability and civility, to facilitating concept formation and growth of understanding. Expandability means that the scientist can define new technical terms and modify existing terms as he creatively peruses the data base. Extensibility is the ability to introduce new data structures by readily modifying the language itself.

Papers XI and XII deal with the PEEP/DECIDE/GRAPH programming system and some elementary applications.

7. ACKNOWLEDGEMENTS

This work was carried out at the University of Pennsylvania, The National Institutes of Health, U.S.A. and the Department of Clinical Cytology, University Hospital of Uppsala. The effort was made possible through support from contracts and grants from the National Institutes of Health, U.S.A. (specifically the National Cancer Institute and the National Bladder Cancer Task Force), and Swedish travel grants obtained through the Department of Clinical Cytology, University Hospital of Uppsala.

Financial support was also given by the Swedish Natural Science Research Council, NO 3701-1 and from the United States National Cancer Institute, contract NO1-CB-53968.

I want to express my gratitude to Docent Björn Stenkvist, head of the Department of Clinical Cytology, University Hospital, Uppsala for his support and enthusiasm.

I further want to thank Professor Heinz Kreiss and Professor Bertil Gustavson of the Department of Computer Science for their support.

I wish to express appreciation to two colleagues: Dr. Mortimer L. Mendelsohn who initiated the leukocyte project and chromosome project, collaborated in the discrimination research, and encouraged the development of new analytical techniques; and Dr. Brian H. Mayall, who expertly prepared, selected and scanned the specimens for optical discrimination.

I also wish to thank my co-author and friend Jan Holmquist for his contribution to the breast cancer project during his pleasant stay at the Division of Computer Research and Technology, National Institutes of Health, U.S.A.

I wish to thank Ms. Stella Barry for skillful assistance with the manuscript.

Last, but not least, I wish to thank my son David Prewitt for his understanding and encouragement.

Bethesda, Maryland in April 1978.

Judith M.S. Prewitt

8. LIST OF PAPERS

This thesis is based on the following papers:

(I) Prewitt, J.M.S.:
The Selection of Sampling Rate for Digital Scanning. IEEE Transaction on Biomedical Engineering, 12: 14-21, 1965.

(II) Mendelsohn, M.L., Mayall, B.H., Prewitt, J.M.S., Bostrum, R.C. and Holcomb, W.G.:
Digital Transformation and Computer Analysis of Microscopic Images. In Advances in Optical and Electron Microscopy, 2 (Barer, R. and Coslett, V., eds.), Academic Press: 77-150, 1968.

(III) Prewitt, J.M.S. and Mendelsohn, M.L.:
The Analysis of Cell Images. Annals of the New York Academy of Sciences 128:1035-1053, 1965.

(IV) Prewitt, J.M.S.:
Parametric and Nonparametric Recognition by Computer: An Application to Leukocyte Image Processing. In Advances in Computers 12 (M. Rubinoff, Ed.), Academic Press:285-414, 1972.

(V) Prewitt, J.M.S., Barber, A.M., Holmquist, J. and Stenkvist, B.:
Computer Assisted Identification of Cells in Needle Aspirates of Mammary Tumors. In Proceedings of Medinfo (D.B. Shires and H. Wolf, eds.), North Holland Publishing Company:693-699, 1977.

(VI) Prewitt, J.M.S.:
A Versatile Clustering Algorithm with Objective Function and Objective Measure. Computer Programs in Biomedicine 2:297-314, 1972.

(VII) Mendelsohn, M.L., Mayall, B.H. and Prewitt, J.M.S.:
Approaches to the Automation of Chromosome Analysis. In Image Processing in Biological Sciences (D. Ramsay, ed.), University of California Press:119-136, 1969.

(VIII) Mendelsohn, M.L., Conway, T.J., Hungerford, D.A., Kolman, W.A., Perry, B.H. and Prewitt, J.M.S.:
Computer-oriented Analysis of Human Chromosomes, I. Photometric Estimation of DNA Content. Cytogenetics 5:223-242, 1966.

(IX) Mendelsohn, M.L., Hungerford, D.A., Mayall, B.H., Perry, B., Conway, T. and Prewitt, J.M.S.: Computer-oriented Analysis of Human Chromosomes, II. Integrated Optical Density as a Single Parameter for Karyotype Analysis. Annals of the New York Academy of Sciences 157:376-393, 1969.

(X) Prewitt, J.M.S.: An Application of Pattern Recognition to Histology. To be published in the Proceedings of IEEE Conference on Image Processing and Pattern Recognition, June, 1978.

(XI) Prewitt, J.M.S., Reece, D.K., Hutchinson, G. and Jackson, C.K.: DECIDE: An Expandable System for Medical Decision-making. Informatique Medicale - Toulouse, 1974, Tome I: Textes des Communications, Institut de Recherche d'Informatique et d'Automatique:153-182, 1974.

(XII) Prewitt, J.M.S.: Interactive Decision-making for Picture Processing. Proceedings of the 1977 IEEE Conference on Decision and Control: 373-379, 1977.

Länstryckeriet - Nyköping 1978